Unveiling the ChatGPT phenomenon: Evaluating the consistency and accuracy of endodontic question answers.

Abstract

AIM

Chatbot Generative Pre-trained Transformer (ChatGPT) is a generative artificial intelligence (AI) software based on large language models (LLMs), designed to simulate human conversations and generate novel content based on the training data it has been exposed to. The aim of this study was to evaluate the consistency and accuracy of ChatGPT-generated answers to clinical questions in endodontics, compared to answers provided by human experts.

METHODOLOGY

Ninety-one dichotomous (yes/no) questions were designed and categorized into three levels of difficulty. Twenty questions were randomly selected from each difficulty level. Sixty answers were generated by ChatGPT for each question. Two endodontic experts independently answered the 60 questions. Statistical analysis was performed using the SPSS program to calculate the consistency and accuracy of the answers generated by ChatGPT compared to the experts. Confidence intervals (95%) and standard deviations were used to estimate variability.

RESULTS

The answers generated by ChatGPT showed high consistency (85.44%). No significant differences in consistency were found based on question difficulty. In terms of answer accuracy, ChatGPT achieved an average accuracy of 57.33%. However, significant differences in accuracy were observed based on question difficulty, with lower accuracy for easier questions.

CONCLUSIONS

Currently, ChatGPT is not capable of replacing dentists in clinical decision-making. As ChatGPT's performance improves through deep learning, it is expected to become more useful and effective in the field of endodontics. However, careful attention and ongoing evaluation are needed to ensure its accuracy, reliability and safety in endodontics.